Block Explanation

1. How did you develop the final blocker? What blocker did you start with? What problems did you see? Then how did you revise it to come up with the next blocker? In short, explain the *development process*, from the first blocker all the way to the final blocker (that you submit in the Jupyter file).

   a. We used the overlap blocker provided by Magellan. In the initial stages of blocking, we constructed a list of stopwords to avoid matches based on common words such as “Ale” or “Brewery”. We removed these words during the process of pulling tuple information from the html pages. In our first iteration of blocking, we isolated the pairs whose names were exactly the same, and as expected, we found this method to be too restrictive, generating roughly 250 tuple pairs. We then attempted to block by country, which proved to be too generous in its matching, leaving roughly 11 million matched pairs. Finally, we decided to block first by brewery and then by beer name, pulling each pair that shared one common word in their beer name. We found that this produced a reasonable output, roughly 2,300 pairs. Overall, we spent more time resolving the issues in the formatting of our tables rather than with the blocker itself.

2. If you use Magellan, then did you use the debugger? If so, where in the process? And what did you find? Was it useful, in what way? If you do not use Magellan, you can skip this question.

   a. We did not use the debugger. Overall, Magellan was helpful because it provided a structure from which to learn and begin blocking easily.

3. How much time did it take for you to do the whole blocking process?

   a. The whole process took ~6 hours.

4. Report the size of table A, the size of table B, the total number of tuple pairs in the Cartesian product of A and B, and the total number of tuple pairs in the table C.

   a. Size of table A: 4,882
   b. Size of table B: 4,084
   c. Product of A and B:
   d. Total Number of tuple pairs in Table C: 2,269

5. Did you have to do any cleaning or additional information extraction on tables A and B?

   a. Yes. We did not realize that some of our IDs from Stage 1 were not unique, so we resolved this issue. In the crawling process, we checked to see if each ID was unique and generated a new ID for those which were not previously unique. Additionally, we applied a dictionary of stopwords to our crawler so that common words such as “Brewery” or “Ale” would not be included in our final tables.

6. Did you run into any issues using Magellan (such as scalability?). Provide feedback on Magellan. Is there anything you want to see in Magellan (and is not there)? If you do not use Magellan, you can skip this question.

   a. Magellan was relatively slow, so it was difficult to apply new blocking schema and receive results quickly. No issues other than time cost.

7. Any other feedback is appreciated.

   a. The tutorial and Jupyter notebook provided by the instructor were very helpful. The output progress bar was helpful, and the ETA was accurate.